Metadata Exchange without pain: the AGRIS AP to harvest and exchange quality metadata

Irene Onyancha, James Weinheimer, Gauri Salokhe, Stephen Katz and Johannes Keizer
{Irene.Onyancha, James.Weinheimer, Gauri.Salokhe, Stephen.Katz, Johannes.Keizer }@fao.org
Food and Agriculture Organization of the United Nations, Rome, Italy
http://www.fao.org/

Abstract

This paper focuses on the AGRIS Application Profile (AGRIS AP), a standard created specifically to enhance description, exchange and subsequent retrieval of agricultural Document-Like Information Objects (DLIOs). The AGRIS AP provides a minimum interoperability layer through which agricultural information can be described and exchanged. The standard, developed in light of the new AGRIS vision, offers the flexibility to enhance the quality of description of agricultural information resources. The paper discusses the advantages of the AGRIS AP as opposed to the current standards by pointing out its strengths, its possible applications and how it will be further developed in the future.

Keywords: Metadata, AGRIS, Application Profiles, Information retrieval, Semantic standards

1. The Problem of Information Exchange

For hundreds of years, librarians have wanted to exchange cataloguing information, but while it may not seem to be an especially arduous task, it has concealed many obstacles. Not only has it been difficult to actually get people to commit to exchanging their information: to go to the trouble of making that extra catalogue card and send it to the proper place or to export records; there were further problems concerning standards and quality: to ensure a single size of a catalogue card or a single computer format. Even once these issues are solved, the problems refuse to disappear, since issues of quality of content arise: what form of the corporate name should be chosen? Which title do I choose?

All of these problems are still with us today, not least because the traditional solutions have always required the acceptance of *uniform standards*, which means that some unfortunate people must change everything they do and accept someone else's standards. As a result, some institutions are forced to abandon their methods, thereby making the labour of years or decades obsolete, so that they can move confidently into the future.

Today, now that computer capabilities have reached a sufficiently high level, this no longer holds true. We believe that the AGRIS AP, although it does not currently solve all of the problems of information exchange, is a major step in the right direction. Anyone can exchange metadata or cataloguing information, and—best of all—no one needs to change a thing that they do. All they have to do is share in the correct way.

In this paper, we shall attempt to show how this has been achieved in the context of the AGRIS[1] information system[*].

Case Scenario

There are several reasons to exchange metadata information: from enhanced searching, to workflow issues, such as avoiding retyping information that has already been input once, or metadata harvesting for value added services. The reasons are many, but what are the problems in exchanging metadata information?

Let's look at a real example of metadata records to see exactly how they differ and what are the consequences of exchanging records. Figures 1 and 2 show two records that describe the same item. Figure 1 is in MARC21/AACR2[2] and Figure 2 is in AGRIN3/AGRIS[3] format.

1.1. Different record structures and applications

The MARC21 record could have been created in many applications (Voyager, ISIS, Horizon, etc.),

[*] AGRIS is the international information system for the agricultural sciences and technology, created by the Food and Agriculture Organization of the United Nations (FAO) in 1974. The main purpose of the AGRIS system is to facilitate information exchange and to bring together scientific and technical literature, especially non-conventional (grey) literature, dealing with all aspects of agriculture.

but it must be in ISO2709 record structure. The AGRIN record is created in CDS/ISIS but could be in Tag Text (a text file with a tag number and relevant value, separated by a carriage return), AGRIN2 (an old CDS/ISIS format) and AGRIN3 (a revised CDS/ISIS format, based on the ISO2709 structure. Many databases use relational database structures, completely bypassing ISO2709.

010 __ |**a** 2001023765
020 __ |**a** 0852382847
040 __ |**a** DLC |**c** DLC |**d** DLC
042 __ |**a** pcc
050 00 |**a** SH328 |**b** .W46 2001
082 00 |**a** 333.95/6/153 |**2** 21
100 1_ |**a** Welcomme, R. L.
245 10 |**a** Inland fisheries : |**b** ecology and management / |**c** compiled by R.L. Welcomme.
260 __ |**a** Oxford ; |**a** Malden, MA : |**b** Fishing News Books, |**c** 2001.
300 __ |**a** xix, 358 p. : |**b** ill., maps ; |**c** 25 cm.
504 __ |**a** Includes bibliographical references (p. 332-352).
650 _0 |**a** Fishery management.
650 _0 |**a** Freshwater fishes |**x** Ecology.
710 2_ |**a** Food and Agriculture Organization of the United Nations.

Figure 1. MARC21/AACR2 record

100: Welcomme, R.L.^b(comp.)
200: Inland fisheries: ecology and management
600: English
401: Oxford (United Kingdom)
402: Fishing New Books for FAO
403: 2001
500: 358 p.
610: graphs, tables; Includes bibliography
320: 0-85238-284-7
800: INLAND FISHERIES; FRESHWATER ECOLOGY; FISHERY MANAGEMENT; FISHERY POLICIES; INLAND WATER ENVIRONMENT; FRESHWATER FISHES; FISHING METHODS; FISH PROCESSING;EVALUATION; GEOGRAPHICAL INFORMATION SYSTEMS;

Figure 2. AGRIN3/AGRIS Record

1.2. Different content designations for the same bibliographic concept

ISO2709 is not enough, but just the beginning of exchange; now we have the problem of different content designations. The MARC21 record is obviously in MARC21, but there are many other MARC formats in the world. Different content designators, where different codes are used to represent the same concept, inhibit interoperability. For example, the MARC21 record uses the 245 field for Title while AGRIN3 uses the 200 field. Another example, 650 in MARC21 and 800 in AGRIN3 represent the same concept (subject) but use different content designators.

New standards have placed new demands for interoperability, for example, the OpenURL[4] standard is not supported by many older content designations.

1.3. Different conceptual bibliographic metadata

The MARC21 record has some concepts that do not exist in the other record. One example is the concept of Main Entry, represented here by the author R.L. Welcomme, who is considered to be the primary author. The AGRIS record does not have the concept of Main Entry.

There are concepts from other standards that neither side completely fulfils: full compliancy with the OpenURL standard, new concepts from the Functional Requirements for Bibliographic Records[5] (FRBR).

1.4. Different cataloguing rules

There are many cataloguing rules in use in the world. The MARC21 format primarily uses the AACR2 cataloguing rules. The AGRIN3 format uses the rules according to the AGRIS Cataloguing guidelines. For example, the place of publication in the AGRIS cataloguing guidelines, is entered as "City (Name of the Country)" while in AACR2, the place of publication is transcribed as it is found on the resource and an additional place of publication is added into the record.
AGRIS: 401: Oxford (United Kingdom)
AACR2: 260__ |a Oxford ; |a Malden, MA

1.5. Variant treatments for different formats (one record/multiple records)

When a similar item appears on the internet in a different format from the printed version, how is it handled: simply by adding the URL to the original record, or is an entirely new record created? In the AGRIN3/AGRIS record, the URL is merely added to the record, whereas in MARC21/AACR2, a new record would most probably be required.

1.6. Multilinguality

Some formats have a greater focus on multiple languages. The AGRIS record shown here has special fields for titles in each language. Therefore, an English title is coded differently from a Spanish title. This is absent in MARC21. There are several

standards for encoding non-ASCII[6] character sets, such as Unicode[7].

1.7. Other Differences in Bibliographic Concepts and Granularity.

Different systems use their own choices of metadata elements which results in different levels of granularity. For example, in MARC21/AACR2, titles are encoded in the following way:

245 10 ^a Inland fisheries : ^b ecology and management

Where:

245=Title statement

10 =Main entry/added entry indicator and filing indicator

^a Title proper

^b Other title information (or subtitle, separated by a space-colon-space)

In AGRIN3/AGRIS format (see Figure 2), this title is encoded as:

200 Inland fisheries: ecology and management

Where:

200= English language title

The above example shows that, depending on the catalogue, the same title of the resource may be added in either one or two different fields.

If the title were in French, it would be placed into a 202 field in AGRIN3/AGRIS record. In MARC21/AACR2, language of Title is irrelevant.

2. Motivation behind the need for a change

The AGRIS New[8] vision is a strategy that was agreed upon by member. It focuses on improving electronic publishing of documentation through continual improvement of web-enabled AGRIS methodologies and tools (with a focus on the establishment of standards), aimed at effective exchange and retrieval of multilingual scientific and technical information.

Although the AGRIS vision is to focus on improving accessibility of science and technology information about agricultural development, its immediate implementation was hampered by some of the existing problems, which have been outlined in section 1 above.

One solution would be to make separate mappings to and from each metadata format, but it turns out that this simply compounds the problem, as illustrated in Figure 3.

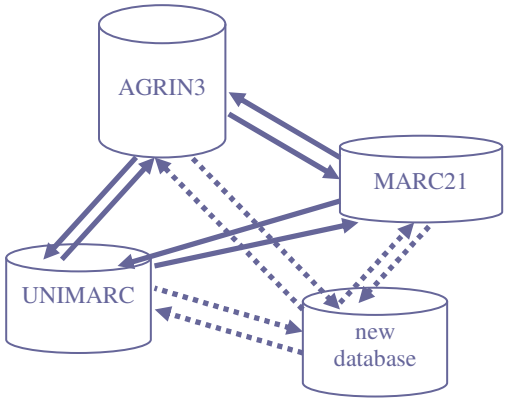

Figure 3. Mappings necessary for sharing data between three or four databases

In the above Figure, the addition of a 'new database' means six new mappings (shown using the dotted lines) will have to be created for everyone to share information with each other. Imagine now, if we had one more! Each new addition of database results in other new mappings. This is calculated with the following formula.

$$n\_P\_2 = \frac{n!}{(n-2)!}$$

Where n = number of databases that want to share information with each other.

Additionally, if there would be a change in any one of the formats, all the other databases that are sharing information would also have to change their own mappings.

3. AGRIS AP: What is it?

Taking into consideration all of these issues, it was clear that we needed new strategies to entice the AGRIS centres to send us data in a platform independent format. The quest for a single standard to describe agricultural resources led us to the conclusion that there would not be a single set that could be used 'as is'. Nevertheless, in order to not reinvent the wheel, we wanted to make use of what is already around and create extensions only where it was absolutely necessary. We needed to define our own application specific profile. Application profiles (or APs) provide the possibility to 'mix and match' what already exists[9]. The AGRIS AP was thus created by taking elements and refinements that are already in existence, such as those declared by organizations like DC and AGLS and those declared by AgMES.

The AGRIS AP is a standard created specifically to enhance description, exchange and subsequent retrieval of agricultural Document-Like Information Objects (DLIOs)[10]. It is a format to produce high quality metadata and allows for a platform-independent exchange of information about different types of agricultural resources. It prescribes a data model by taking specific elements from established namespaces† namely; Agricultural Metadata Element Set (AgMES)11 and the Dublin Core Metadata Element Set (DCMES)12 and Australian Government Locator Service (AGLS)13. The use of well accepted standards improves both interoperability and resource discovery, and at the same time it promotes reuse and restricts reinvention. AgMES was created to accommodate elements, refinements and schemes that are necessary for description and discovery of agricultural information resources. It does not reinvent the elements, but only extends the DC where necessary. These extensions considerably improve the quality of metadata and subsequently improve the access and retrieval of the information.

The AGRIS AP consists of 15 core elements, 43 refinements and 32 schemes and provides best recommended practices for entry of data on each of the elements and refinements. It also provides information on cardinality, obligation and the allowed data format.

4. Benefits of the AGRIS AP

The AGRIS Application profile as an exchange format addresses the significant aspects of metadata interoperability. It:

- reuses content designators recommended by Dublin Core. It also uses elements that have been declared in other standards such as AGMES and AGLS
- uses XML and RDF syntax for coding. These syntaxes and widely applied for exchange and storage of information.
- is both human and machine readable.

This interoperability allow for various value-added services.

1. **Exchange of agricultural information.** Many partners, especially those from developed countries, are now contributing to the AGRIS central database using the AGRIS AP. The AGRIS AP enables exchange of data from systems that are using cataloguing and management rules other than those prescribed by the AGRIS guidelines. For example, the Netherlands AGRIS Resource centre uses a local format for management of its resource; however, it has submitted data using the AGRIS AP for the AGRIS central database.
2. **Harvesting of metadata and associated content for Open Archives.** The generic level of the AGRIS AP compliments the unqualified Dublin Core metadata set. This is the recommended format used for metadata harvesting in the Open Archives Initiative[14]. The AGRIS AP facilitates exposure of the AGRIS content to the OAI systems, making it harvestable and available to a wider audience.
3. **Possibility to access the actual resource from the Web.** A recent study concluded that more and more resources are being retrospectively added to the Web. Based on this study, availability of good quality metadata allows for retrieval of the original resource, regardless of its actual location on the Web[15].

Additionally, the XML format combined with XSL Transformations can allow for information to be used in many unique ways. This is potentially limitless. For example, a user could take a record encoded in XML, place it into a special XSL style sheet, and format a perfectly styled bibliographical reference in a word processing program.

4.1. AGRIS AP exchange layer avoids the need for a single format

Mapping all of the different formats (MARC21, UNIMARC, FINMARC, RUSMARC, JPMARC, AGRIN to name just a few) has proven to be practically impossible, especially since changes to formats occur with some regularity. As a result, there has been tremendous pressure for everyone to accept a single MARC format, but this also involves no less work to convert entire catalogues—except for those lucky few who already happen to use the chosen format.

Another suggestion is to use an “exchange layer” that would serve for exchanging all bibliographic information, whether it is in one of the versions of MARC ISO2709, or any other, perhaps relational database, structure. Therefore, if you could put in one field, you would receive the

† An XML namespace is a collection of names, identified by a URI reference which are used in XML documents as element types and attribute names. XML namespaces differ from the "namespaces" conventionally used in computing disciplines in that the XML version has an internal structure and is not, mathematically speaking, a set. (Taken from http://www.w3.org/TR/REC-xml-names/.)

corresponding field for the other formats. [See Figure 4]

For example: in MARC21, information for the "Publisher" goes into field 260, subfield b; in UNIMARC, it goes into field 210, subfield c; in AGRIN3 (used by FAO), it is placed in field 402.

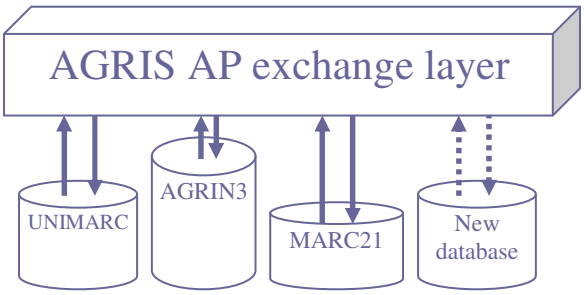

Figure 4: Mapping different formats to the AGRIS AP format

The advantage of the situation illustrated above is that each format needs only to create conversions to and from the exchange layer and avoid the need to create separate conversions to and from all the other formats. Any changes within one format would not result in reciprocal changes for everyone else, since the changes would affect only their own output/input to the exchange layer. If this worked for all fields and all formats, any record could be shared with any database. Of course, local editing would still be necessary in many cases, but no one would need to change anything within their own local databases. The AGRIS AP attempts to be this "exchange layer".

4.2. Platform independent exchange to facilitate interoperability and reusability of information

The method to achieve simple metadata exchange is through the use common metadata standards and a common syntax, the XML format. The use of both these aspects in the AGRIS AP enhances the possibility for exchange avoiding many of the earlier problems related to systems. The result now is that the format or structure of any database is irrelevant for metadata exchange.

4.3. Supporting multilinguality

The xml:lang=" " attribute is used for elements for which it was considered necessary to know the language of its content. This extensibility enables multiple values of the specified field in any language. It was already mentioned that titles were indicated as being only in English, Spanish, French or Other. However, with the new specifications, title element can be in any language as long as the language is indicated using the xml:lang attribute.

Example: Titles element provided in Dutch (nl) and English (en).

```
<dc:title xml:lang="nl"> Waterwijs : plannen
met water op regionale schaal
    <dcterms:alternative xml:lang="en">
    [Carefull with water: plans with water on
    regional scale] </dcterms:alternative>
</dc:title>
```

On the structural side, there are different methods of inputting non-ASCII character sets, including Windows character sets, DOS, ISO, along with special bibliographical character sets, as exists in MARC-8[16]. Unicode is now widely accepted as a standard and has been implemented in many systems; AGRIS-AP has chosen to accept UTF-8[17] character encodings.

4.4. Maintain a level of quality in the collected information

Certain elements are critical for searching and necessitate the use of metadata schemes, thesauri and controlled lists. The use of these schemes also assures a level of consistency to be achieved in the collected information. The AGRIS AP recognises such elements and provides controlled vocabularies, lists and Schemes for these elements.

The following example shows how the AGRIS AP offers a means by which different controlled vocabularies and recommended schemes in agricultural sciences and technology could be used. The schemes that are used for the subject element are specific to the Agricultural Community. They provide the source information and thus the possibility of providing value-added searches.

The following subjects have all been given to the same item: the first is AGROVOC[18], second is CABI Thesaurus[19], and third is Library of Congress Subject Headings[20].

Example: Subject metadata encoded with codes from AGROVOC, CAB Thesaurus and LCSH

```
<dc:subject>
    <ags:subjectThesaurus xml:lang="en"
    scheme="ags:AGROVOC">Animal
    Husbandry </ags:subjectThesaurus>
    <ags:subjectThesaurus xml:lang="en"
    scheme="ags:AGROVOC">Livestock
    Management </ags:subjectThesaurus>
    <ags:subjectThesaurus xml:lang="en"
    scheme="ags:AGROVOC">Animal
    Research</ags:subjectThesaurus>
    <ags:subjectThesaurus xml:lang="en"
    scheme="ags:CABT">age-
    differences</ags:subjectThesaurus>
    <ags:subjectThesaurus xml:lang="en"
    scheme="ags:CABT">animal-
    husbandry</ags:subjectThesaurus>
    <ags:subjectThesaurus xml:lang="en"
    scheme="ags:LCSH">Livestock systems--
    Congresses </ags:subjectThesaurus>
</dc:subject>
```

With minor changes to the DTD, AGRIS could also accept records catalogued with keywords from other thesauri as well. Therefore, a record catalogued with keywords from the Bibliotheque National de France and the Swiss National Library would be validated with the modified AGRIS DTD.

Example: Subject metadata encoded with codes from RAMEAU and SWD

```
<dc:subject>
    <ags:subjectThesaurus xml:lang="fr"
    scheme="ags:RAMEAU">Bétail --
    Alimentation </ags:subjectThesaurus>
    <ags:subjectThesaurus xml:lang="de"
    scheme="ags:SWD">Nutztierhaltung
    </ags:subjectThesaurus>
    <ags:subjectThesaurus xml:lang="de"
    scheme="ags:SWD">Haltungssystem
    </ags:subjectThesaurus>
</dc:subject>
```

Thus, all the individual terms can still be used for searching, but each one can also be "fine-tuned" to search only with terms from RAMEAU[21] or AGROVOC, if desired.

4.5. New possibilities

The AGRIS AP's structure and content plays a major role in enhanced searching and retrieval of agricultural documents. The XML structure, in which the metadata is encoded, significantly improves the usefulness of the data. AGRIS data, exported using the AGRIS AP, has been reused in different scenarios to achieve value added services.

Using the AGRIS AP has the following advantages that support searching and retrieval.

- The ability to search multiple databases with a single search has long been possible through the Z39.50 protocol[22], but has not been widely implemented because of limitations on searching and record display. Using AGRIS-AP and web services[23] offers possibilities that are far greater than ever before. Web services allow for superior searching capabilities, while XSL Transformations[24] can sort records, eliminate duplicates and do further processing of the records received by the user.
- OpenURL[25] and SICI [26] are just a couple of the newer standards that have proven to be highly successful in assisting search and retrieval of journal articles. Metadata records must be interoperable with these standards and others that may arise.
- RSS newsfeeds[27] have become very popular and are an excellent example of reusing XML applications that were originally designed for non-bibliographic uses. Installing an RSS feeder is relatively easy, and can be used to notify users of new items in their own, specific interests.

5. Future Issues

Currently, there are several limitations of the AGRIS-AP that make it difficult to use it as a common exchange standard. Formats and cataloguing rules tend to work together, e.g. MARC21 and AACR2 mirror one another. AGRIS-AP is based on the rules of the AGRIS cataloguing system, which is non-ISBD. ISBD[28] serves as the foundation for the majority of cataloguing codes used by the national libraries and major bibliographic agencies.) It lacks several bits of information of critical importance for ISBD-based cataloguing rules: e.g. no statement of responsibility, different rules for the extent statement and so on. For greater harmony, AGRIS-AP must be enhanced to allow for greater record sharing of ISBD-based records. In other cases, bibliographic treatments are not completely the

same: FAO-managed development projects are treated in a special way in the AGRIS database, while they are considered as normal corporate bodies in most other databases.

In spite of our best efforts however, it should now be clear that loss of data is inevitable when exchanging metadata information. This is because there are concepts of bibliographic entities that are not shared by both sides: in the one case, encoding the language for the title, in the other case, no main/added entry or filing indicators. There is also the issue of granularity where one assigns separate codes for title and subtitle, while the other puts it into a single field. This also leads to loss of data.

The AGRIS-AP currently deals with the levels of structure and content designations. However, the content itself is very difficult to standardize and will remain an ongoing problem.

This may be helped as the Functional Requirements for Bibliographic Records are implemented (and modified) by bibliographic agencies around the world. AGRIS must also do its best to interoperate with these requirements. Some elements used by FRBR do not exist in AGRIS AP, for example, uniform title. Overcoming these problems will be difficult.

6. Conclusion

The AGRIS AP offers strong motivations why it should be adopted as a standard for description and exchange of agricultural resources. Due to its simplicity, the application has functions that the standards presently available do not offer. It offers both a generic format that is suitable for exchange of information at a minimal level and a richer format that supports higher quality metadata. Therefore, it is being recommended as a standard for exchange of metadata about DLIOs in the agricultural community.

Acknowledgements

The authors wish to acknowledge the assistance of Anita Liang, Stefka Kaloyanova and Stefano Anibaldi in the preparation of this article.

References

[1] AGRIS home page http://www.fao.org/agris/

[2] MARC Standards / Library of Congress. http://www.loc.gov/marc/ . See also: Anglo-American Cataloguing Rules: 2002 Revision. American Library Association

[3] AGRIS: guidelines for bibliographic description and input sheet preparation. Rome: Food and Agriculture Organization of the United Nations, Jan. 1998. Available from: ftp://ext-ftp.fao.org/GI/agris/pdf/guidelns/main.pdf

4 The OpenURL standard. http://www.niso.org/committees/committee_ax.html

[5] Functional Requirements for Bibliographic Records. http://www.ifla.org/VII/s13/frbr/frbr.pdf

6 Non Ascii character set.

7 Unicode standard. http://www.unicode.org/

[8] AGRIS - A strategy for an international network for information in agricultural sciences and technology within the WAICENT Framework http://www.fao.org/docrep/MEETING/005/AC502e.htm

[9] Heery, Rachel and Manjula Patel (2000) "Application profiles: mixing and matching metadata schemas". Ariadne, No. 25, September. http://www.ariadne.ac.uk/issue25/app-profiles/intro.html

[10] AGRIS AP Manual. ftp://ext-ftp.fao.org/agris/agmes/AGRISAP-UserGuide.doc

[11] Agricultural Metadata Element Set Name Space. ftp://ext-ftp.fao.org/agris/agmes/AGMESNS-DLIO.doc

[12] Dublin Core Metadata Initiative. http://dublincore.org/

[13] Australian Government Locator Service. http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html

[14] Open Archives Initiative. http://lcweb.loc.gov/cds/lcsh.html#lcsh20

[15] Salokhe, G., Weinheimer, J., Bovo, M.G., Agrimi, M. “Structured Metadata for Direct Resource Location: A Case Study” (2003). http://www.siderean.com/dc2003/404_Paper84-color.pdf

[16] MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media. http://www.loc.gov/marc/specifications/speccharmarc8.html

[17] UTF 8 Standard. http://www.utf-8.com/

[18] AGROVOC Multilingual Agricultural Thesaurus http://www.fao.org/agrovoc/

[19] CAB Thesaurus http://www.cabi-publishing.org/

[20] Library of Congress Subject Headings. http://lcweb.loc.gov/cds/lcsh.html#lcsh20

[21] Bibliotheque National de France, RAMEAU. http://rameau.bnf.fr/informations/index.htm

[22] Z39.50 Maintenance Agency Page. http://www.loc.gov/z3950/agency/

[23] Web Services. http://www.w3.org/2002/ws/
[24] XSL and XSL Transformations. http://www.w3.org/Style/XSL/
[25] The OpenURL Framework for Context-Sensitive Services / NISO. http://www.niso.org/committees/committee_ax.html
[26] SICI : Serial Item and Contribution Identifier Standard. http://sunsite3.berkeley.edu/SICI/
[27] RSS 2.0 Specification http://blogs.law.harvard.edu/tech/rss
[28] International Standard Bibliographic Description. http://www.ifla.org/VII/s13/pubs/isbd.htm